On the acceleration of some empirical means with application to nonparametric regression

Bernard Delyon, François Portier

On the acceleration of some empirical means with application to nonparametric regression

Bernard Delyon and François Portier

Abstract: Let $(X_1, \dots, X_n)$ be an i.i.d. sequence of random variables in $\mathbb{R}^d$, $d \geq 1$, for some function $\varphi : \mathbb{R}^d \to \mathbb{R}$, under regularity conditions, we show that

$$n^{1/2} \left(n^{-1} \sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)} - \int \varphi(x) dx \right) \xrightarrow{\mathbb{P}} 0,$$

where $\widehat{f}^{(i)}$ is the classical leave-one-out kernel estimator of the density of X_1. This result is striking because it speeds up traditional rates, in root n, derived from the central limit theorem when $\widehat{f}^{(i)} = f$. As a consequence, it improves the classical Monte Carlo procedure for integral approximation. The paper mainly addressed with theoretical issues related to the later result (rates of convergence, bandwidth choice, regularity of φ) but also interests some statistical applications dealing with random design regression. In particular, we provide the asymptotic normality of the estimation of the linear functionals of a regression function on which the only requirement is the Hölder regularity. This leads us to a new version of the *average derivative estimator* introduced by Härdle and Stoker in [13] which allows for *dimension reduction* by estimating the *index space* of a regression.

Key words: Semiparametric regression, Multiple index model, Kernel smoothing, Integral approximation.

1 Introduction

Let $(X_1, \dots, X_n)$ be an i.i.d. sequence of random variables in $\mathbb{R}^d$, $d \geq 1$, for some function $\varphi : \mathbb{R}^d \to \mathbb{R}$, under regularity conditions, we show that

$$n^{1/2} \left(n^{-1} \sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)} - \int \varphi(x) dx \right) \xrightarrow{\mathbb{P}} 0, \tag{1}$$

where $\widehat{f}^{(i)}$ is the classical leave-one-out kernel estimator of the density of X_1 say f, defined by

$$\widehat{f}^{(i)}(x) = (nh^d)^{-1} \sum_{j \neq i}^{n} K(h^{-1}(X_j - x)), \qquad \text{for every } x \in \mathbb{R}^d,$$

where K is a d-dimensional kernel and where h, called the bandwidth, needs to be chosen and will certainly depend on n. Result (1) and the central limit theorem lead to the following reasoning: when estimating the integral of a function that is evaluated on a random grid (X_i), whether f is known or not, using a kernel estimator of f provides better convergence rates than using f itself. A first obvious application of this result is for Monte Carlo integration when

the design, i.e. the distribution of the points, is not controlled. If the design is free, other methods exist, like quasi random numbers, which may prove to be more efficient, depending on the regularity of the function and on the dimension (we refer to [1] for a comprehensive presentation of these methods). In this paper, we are interested in the random design case for which the previous methods as Quasi Monte Carlo and grid integration cannot be implemented.

Equation (1) may have applications in nonparametric regression with random design. Let

$$Y_i = g(X_i) + \sigma(X_i)e_i, \tag{2}$$

where (e_i) is an i.i.d. sequence of real random variables independent of the sequence (X_i), and $\sigma : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ are unknown functions. In this context, one of the most evident use of Equation (1) deals with the estimation of the linear functionals of g, i.e. the quantities $\int g(x)\psi(x)dx$ for some functions $\psi : \mathbb{R}^d \to \mathbb{R}$. Under regularity conditions, we show that

$$n^{1/2}\left(n^{-1}\sum_{i=1}^{n}\frac{Y_i\psi(X_i)}{\widehat{f}^{(i)}(X_i)} - \int g(x)\psi(x)dx\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, v), \tag{3}$$

where $v = \mathrm{var}((Y - g(X_1))\psi(X_1)f(X_1)^{-1})$. Among typical applications of Result (3), we can mention Fourier coefficients estimation for either nonparametric estimation (see for instance [12], section 3.3) or location parameter estimation (see [9] and the reference therein). We shall focus on applications dealing with the *multiple index model*, i.e. when the link function $g(x) = g_0(\beta^T x)$ for every $x \in \mathbb{R}^d$, with $\beta \in \mathbb{R}^{d\times p}$ called the index, $p \leq d$. As it was noticed by Härdle and Stoker in [13], for the *average derivative estimator* (ADE), when $\psi = \nabla f$ the estimator in Equation (3) recovers the *index* with rates root n. Their method is popular, notably because it is a direct estimation procedure that does not involve complicated optimization algorithm. Thanks to Result 3, we shall see that choosing different functions ψ than ∇f may lead to an accurate estimation of the *index space* $\mathrm{span}(\beta)$.

The estimation of the linear functionals of g is a typical semiparametric problem in the sense that it requires the nonparametric estimation of f as a first step and then to use it in order to estimate a real parameter. To the best of our knowledge, estimators that achieve root n consistency have not been provided yet in the case of a regression with random design. Our approach is based on kernel estimates $\widehat{f}^{(i)}$ of the density of X_1 that is then plugged into the classical empirical estimator of the quantity $\mathbb{E}[Y\psi(X)f(X)^{-1}]$. There is at least four main interesting facts about the weak convergence (3). They are listed below.

(A) The first point about Equation (3) is that, despite slower rates than root n obtained when estimating f, the final estimator recovers the parametric rate root n. Similar facts have already been noticed by some authors in different semiparametric problems as, among others, by Stone in [23] in the case of the estimation of a location parameter, by Robinson in [21] in a *partially linear regression model*, or by Härdle and Stoker in [13] studying ADE (see also [15] and [5] about the semiparametric M-estimation).

(B) Going further in the analysis of Result (3), we notice that the asymptotic variance v is smaller than the asymptotic variance of the estimator with the true density (see Equation (8) in Remark 7). As a consequence for this problem, there is an asymptotic gain in estimating the density. We might remark that the underlying cause is Result (1) because it implies that the asymptotic variance v stems only from the noise e_i associated to the observation of Y_i in Model (2). Surprisingly there is not any terms in v that are due to the randomness of the design.

(C) Despite similarities between our estimator and some estimators of the semiparametric literature (e.g. the references in Point (A)), the technical details of our approach are different since they are based on Equation (1). A similar result was originally stated by Vial in [26] (Chapter 7, Equation (7.27)) in the *multiple index model* context.

(D) Unfortunately, it turns out that Result (1) is no longer true when estimating functionals of the form $f \mapsto \int T(x, f(x))dx$ where $T : \mathbb{R}^2 \to \mathbb{R}$ is different from the map $(x, y) \mapsto \varphi(x)$ (see Section 5). As a result, it suggests that Point (B) has no reason to hold when estimating $\int g(x)T(x, f(x))dx$ with our approach. In view of the asymptotic variance of ADE expressed in Equation (12), this kind of suboptimal properties happen also for ADE where the transformation T differs from the map $(x, y) \mapsto \varphi(x)$ and involves the derivative of f. As a consequence, it might be better to replace, in ADE, the derivatives of f by the derivatives of a known function.

The paper is organized as follows. Section 2 deals with technical issues related to Equation (1). In particular, we examine the rates of convergence of (1) according to the choice of the bandwidth, the dimension and the regularity of the functions φ and f. We also introduce a corrected estimator that converges to 0 faster than the initial one given in Equation (1). This corrected estimator allows a less restrictive choice of the bandwidth. Section 3 is dedicated to the estimation of the linear functionals of g. We show Result (3) under mild conditions on g that only needs to be piecewise Hölder. In Section 4, we focus on the application of our results in the context of the *multiple index model*. We provide a new version of ADE that might be more efficient (see point (D). We give some simulations that compare our method with ADE and *inverse regression methods* introduced by Li in [17] that typically ask more than ADE on the distribution of X.

2 Integral approximation by kernel smoothing

Let $Q \subset \mathbb{R}^d$ be the support of φ. The quantity $I(\varphi) = \int \varphi(x)dx$ is estimated by

$$\widehat{I}(\varphi) = n^{-1} \sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)}$$

We define the leave-one-out estimator of the variance of $h^{-p}K(h^{-1}(x - X_j))$ by

$$\widehat{v}^{(i)}(x) = ((n-1)(n-2))^{-1} \sum_{j \neq i}^{n} (h^{-d}K(h^{-1}(x - X_j)) - \widehat{f}^{(i)}(x))^2,$$

this one is needed to correct the initial estimator by

$$\widehat{I}_{cor}(\varphi) = n^{-1} \sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}^{(i)}(X_i)} \left(1 - \frac{\widehat{v}^{(i)}(X_i)}{\widehat{f}^{(i)}(X_i)^2}\right).$$

To state our main result about the convergences of $\widehat{I}(\varphi)$ and $\widehat{I}_{cor}(\varphi)$, we define the Nikol'ski class $\mathcal{H}_s$ of functions of regularity $s = k+\alpha$, $k \in \mathbb{N}$, $0 < \alpha \leq 1$ as the set of k times differentiable functions φ such that all its derivatives of order k satisfy [25]

$$\int (\varphi^{(l)}(x + u) - \varphi^{(l)}(x))^2 dx \leq C|u|^{2\alpha}, \quad l = (l_1, \ldots, l_d), \quad \sum l_i \leq k. \tag{4}$$

Be careful that $k = \lfloor s \rfloor$, with the convention that $\lfloor n \rfloor = n - 1$ if $n \in \mathbb{N}$. We need the following assumptions.

(A1) For some $s > 0$ the function φ belongs to $\mathcal{H}_s$ on $\mathbb{R}^d$ and has compact support Q.

(A2) The variable X_1 has a bounded density f on $\mathbb{R}^d$ such that its r-th order derivatives are bounded.

(A3) For every $x \in Q$, $f(x) \geq b > 0$.

(A4) The kernel K is symmetric with order $r \geq s$. Moreover, for every $x \in \mathbb{R}^d$, $K(x) \leq C_1 \exp(-C_2\|x\|)$ for some constants C_1 and C_2.

The next theorem is proved in the appendix.

Theorem 1. *Assume that* (A1-A4) *hold, we have the following* $O_{\mathbb{P}}$ *estimates*

$$n^{1/2}\left(\widehat{I}(\varphi) - \int \varphi(x)dx\right) = O_{\mathbb{P}}\left(h^s + n^{1/2}h^r + n^{-1/2}h^{-d}\right), \tag{5}$$

$$n^{1/2}\left(\widehat{I}_{cor}(\varphi) - \int \varphi(x)dx\right) = O_{\mathbb{P}}\left(h^s + n^{1/2}h^r + n^{-1/2}h^{-d/2} + n^{-1}h^{-3d/2}\right) \tag{6}$$

which are valid if the sums inside the $O_{\mathbb{P}}$*'s tend to zero.*

Remark 1. Assumption (A2) about the smoothness of f is crucial to guarantee the convergences stated in Theorem 1. On the one hand, r needs to be greater than d to obtain convergence (5), on the other hand, r greater than $3d/4$ suffices to get convergence (6). In the case where each previous assumption fails, there does not exist h such that Equation (5) or Equation (6) hold. This phenomenon is often referred as the *curse of dimensionality*. The choice of the bandwidth can be made regarding the $O_{\mathbb{P}}$ estimates in Theorem 1 and assuming that $h = Cn^{-a}$. To select the parameter a, one can optimize the quantity in the $O_{\mathbb{P}}$ in order to derive the best possible rate of convergence. For instance, assuming that r and s are sufficiently large so that the first terms in equations (5) and (6) and the last term in Equation (6) are negligible ($2r > 3d$ and $2s > r - d/2$), we obtain the optimal rates $n^{-\frac{(r-d)}{2(r+d)}}$ and $n^{-\frac{(r-d/2)}{2(r+d/2)}}$ for bandwidth $h \propto n^{-\frac{1}{r+d}}$ and $h \propto n^{-\frac{1}{r+d/2}}$, respectively. As in the semiparametric problem studied in [13] (see section 4.1), our estimator of f is suboptimal with respect to the density estimation problem (see [24]). Indeed, to achieve the optimal rates of density estimation one needs to have $h \propto n^{-1/(2r+d)}$ which contradicts the fact that the bias goes to 0 in Theorem 1. However the choice of the constant C in the bandwidth is not studied here. One can follow Härdle, Hart, Marron and Tsybakov (1992) and optimized an equivalent of the MSE, in order to obtain C.

Remark 2. Assumption (A2) neglects the bias problems in the estimation of f that may occur at the borders of Q. Indeed, if f has a jump on the boundary of Q, then our estimate of f would be asymptotically biased and the rates provided by Theorem 1 does not hold. To get ride of this problem, one can correct by hand the estimator, as for instance in [16], or use Beta kernels as detailed in [2]. The simulations provided in Figure 1 highlight how this problem affects the estimation by considering two different densities.

Remark 3. Assumption (A3) basically says that f is separated from 0 on Q. The exponential bound on the kernel in Assumption (A4) guarantee that f is estimated uniformly on Q (see [6]). This leads to $(\inf_{x\in Q} \widehat{f}(x))^{-1} = O_{\mathbb{P}}(1)$ and helps to control the random denominator $\widehat{f}^{(i)}(X_i)$ in the expression of $\widehat{I}(\varphi)$ and $\widehat{I}_{cor}(\varphi)$. In the context of Monte Carlo procedure for integral approximation, Assumption (A2) and Assumption (A3) are not at all restrictive because it is always possible to draw the X_i's from any probability distribution smooth enough and whose support contains the integration domain.

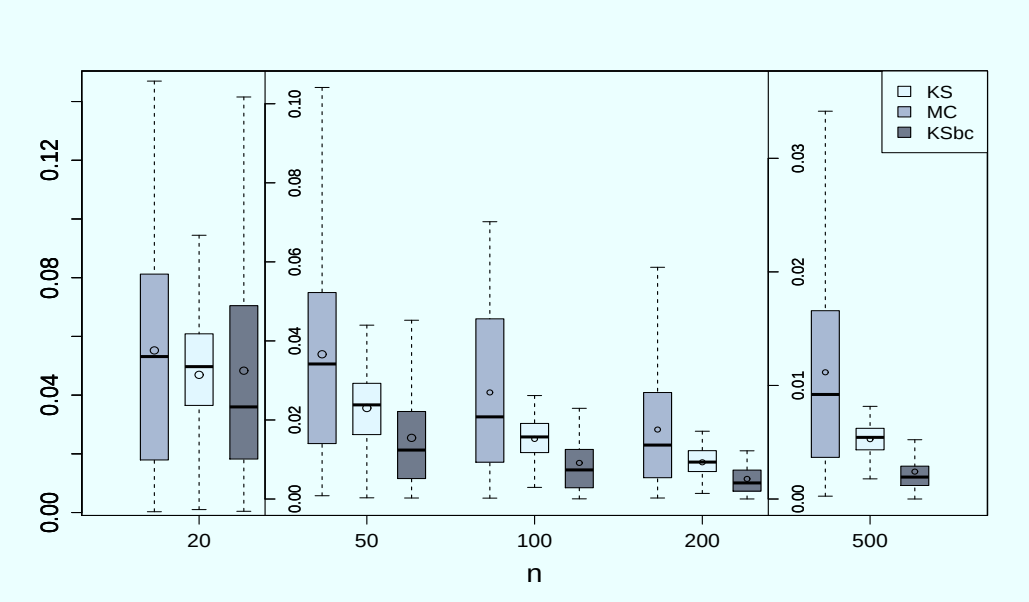

Figure 1: Boxplot over 100 samples of the estimation error of $\int_0^1 \sin(\pi x)dx$, by the classical Monte Carlo procedure with $f = 1_{[0,1]}$, noted MC; by the kernel smoothing with $f = 1_{[0,1]}$, the Epanechnikov kernel and $h = n^{-1/3}$, noted KS; by the kernel smoothing with bias correction with $f = 1_{[-h,1+h]}$ the Epanechnikov kernel and $h = n^{-1/3}$ noted KSbc; for different sample number.

Remark 4. The use of leave-one-out estimators in $\widehat{I}(\varphi)$ and $\widehat{I}_{cor}(\varphi)$ is not only justified by the simplification they involve in the proof (some diagonal terms disappear from the sums). It also leads to better convergence rates. For instance, let us consider the term $\widehat{R}_0$ in the proof of Equation (6) in Theorem 1. Replacing the leave-one-out estimator of f by the classical one, $\widehat{R}_0$ remains a degenerate U-statistic but with nonzero diagonal terms. It is easy to verify that those terms lead to the rates $n^{-1/2}h^{-d}$) which is greater than the rate we found for $\widehat{I}_{cor}(\varphi)$.

Remark 5. The function class $\mathcal{H}_s$ contains two interesting sets of functions that provide different rates of convergence in Theorem 1. First, if φ is α-Hölder on $\mathbb{R}^p$ with bounded support, then φ belongs to $\mathcal{H}_\alpha$. Secondly, if the support of φ is a bounded convex set and φ is α-Hölder inside its support (e.g. the indicator of a ball) then $\varphi \in \mathcal{H}_{\min(\alpha,1/2)}$ (see Theorem 6 in the appendix). As a result, this loss of smoothness at the boundary of the support involves a loss in the rates of convergence (5) and (6). Precisely, whatever the smoothness degree of the function inside its support, if continuity fails at the boundary, rates are at most in $h^{1/2}$.

3 Estimating the linear functionals of a regression function

Let $Q \subset \mathbb{R}^d$ be a compact set and $L_2(Q)$ be the space of squared-integrable functions on Q. We endowed $L_2(Q)$ with the canonical inner product so that it is an Hilbert space. We consider model (2) assuming that $g \in L_2(Q)$. Let $\psi \in L_2(Q)$ be extended to $\mathbb{R}^d$ by 0 outside of Q (ψ has compact support Q). The inner product in $L_2(Q)$ between the regression function g and ψ, is given by

$$c = \int g(x)\psi(x)dx,$$

note that if ψ belongs to a given basis of $L_2(Q)$, then c is a coordinate of g inside this basis. We define the estimator

$$\widehat{c} = n^{-1} \sum_{i=1}^{n} \frac{Y_i \psi(X_i)}{\widehat{f}^{(i)}(X_i)},$$

to derive the asymptotic of $\sqrt{n}(\widehat{c} - c)$, we use Model (2) to get the decomposition

$$\sqrt{n}(\widehat{c} - c) = \widehat{S} + \widehat{R}, \tag{7}$$

with

$$\widehat{R} = \ n^{-1/2} \left(\sum_{i=1}^{n} \frac{g(X_i)\psi(X_i)}{\widehat{f}^{(i)}(X_i)} - \int g(x)\psi(x)dx \right)$$
$$\widehat{S} = n^{-1/2} \sum_{i=1}^{n} \frac{\sigma(X_i)\psi(X_i)}{\widehat{f}^{(i)}(X_i)} e_i.$$

Under some conditions, Theorem 1 provides that $\widehat{R}$ is negligible with respect to $\widehat{S}$. As a result, $\widehat{S}$ carries the weak convergence of $\sqrt{n}(\widehat{c} - c)$, and then the limiting distribution can be obtained making full use of the independence between the X_i's and the e_i's. In order to follow this program, this assumptions are needed.

(A5) The function ψ is Hölder on its support $Q \subset \mathbb{R}^d$ nonempty bounded and convex.

(A6) The function g is Hölder on Q and σ is bounded.

(A7) The bandwidth verifies $n^{1/2}h^r \to 0$ and $n^{1/2}h^d \to +\infty$ as n goes to infinity.

The following theorem is proved in the appendix.

Theorem 2. *Assume that* (A2-A7) *hold, we have*

$$n^{1/2}(\widehat{c} - c) \xrightarrow{\mathrm{d}} \mathcal{N}(0, v),$$

where v is the variance of the random variable $\frac{Y_1 - g(X_1)}{f(X_1)}\psi(X_1)$.

Remark 6. The set Q reflects the domain where g is studied. Obviously, the more dense the X_i's in Q, the more stable the estimation. Nevertheless, it could happened that f vanishes somewhere on Q and this is not taken into account by our framework. In such situations, one may adapt Q from the sample such that the estimated density does not take too small values. This method called *trimming* (employed for instance in [13]) guarantees computational stability as well as some theoretical properties. Even if such an approach is feasible here, it involves much more technicalities in the proofs and may cause a loss in the clarity of the statements.

Remark 7. The nonstandard convergence rates observed in Theorem 1 impacts Theorem 2 in the following way. Let us compare both estimate $\widehat{c}$ and $\widetilde{c} = n^{-1} \sum_{i=1}^{n} Y_i \psi(X_i) f(X_i)^{-1}$ where the latter requires to know f. First, if the signal is observed without noise, that is $Y_i = g(X_i)$, then $n^{1/2}(\widehat{c} - c)$ goes to 0 in probability and $\widetilde{c}$ is asymptotically normal. Secondly, when there is some noise in the observed signal, that is $e_i \neq 0$, the comparison can be made regarding their asymptotic variances. Since we have

$$v = \mathrm{var}\left(\frac{Y_1}{f(X_1)} \psi(X_1) \right) - \mathrm{var}\left(\frac{g(X_1)}{f(X_1)} \psi(X_1) \right) \leq \mathrm{var}(n^{1/2}(\widetilde{c} - c)), \tag{8}$$

it is asymptotically better to plug the nonparametric estimator of f than to use f directly.

4 Applications to multiple index models

4.1 Average derivative estimator

The multiple index model is defined as Model (2) with the specification

$$g(x) = g_0(\beta^T x), \qquad \text{for every } x \in \mathbb{R}^d, \tag{9}$$

where $\beta \in \mathbb{R}^{d\times p}$, and p is minimal. Under some conditions, essentially that X_1 has a density [18], $E = \text{span}(\beta)$ is unique, it is called the *index space* and the term *index* denotes any of its basis. From now, we assume that E is unique. Our approach is based on the gradient of the regression curve since $\nabla g(x) \in E$.

Under some regularity conditions (see [20]), by the integration by parts formula, we have that

$$\beta_\psi = \int g(x)\nabla\psi(x)dx = -\int \nabla g(x)\psi(x)dx \in E, \tag{10}$$

for any smooth function $\psi : \mathbb{R}^d \to \mathbb{R}$. In view of Theorem 2, the following estimator

$$\widehat{\beta}_\psi = n^{-1}\sum_{i=1}^{n} \frac{Y_i \nabla\psi(X_i)}{\widehat{f}^{(i)}(X_i)}, \tag{11}$$

is root n-consistent in estimating a direction of the index space. By applying Theorem 2, we obtain the following corollary where (A5) becomes

(A5) The function $\nabla\psi$ is Hölder on its support $Q \subset \mathbb{R}^d$ nonempty bounded and convex.

Corollary 3. *Assume that* (A2-A7) *hold, we have*

$$n^{1/2}(\widehat{\beta}_\psi - \beta_\psi) \xrightarrow{\text{d}} \mathcal{N}(0, v),$$

where v is the variance of the random variable $\frac{Y_1 - g(X_1)}{f(X_1)}\nabla\psi(X_1)$.

In order to recover the whole space E, we have to compute several $\widehat{\beta}_\psi$, say $(\widehat{\beta}_1, \cdots, \widehat{\beta}_K)$ associated with several functions $\psi = \psi_1, \ldots \psi_K$ and assume in addition that Equation (10) holds true for each $\widehat{\beta}_k$. The estimate $\widehat{E}$ of E will be taken as the p-dimensional space from which the $\widehat{\beta}_k$'s are the closest; there is several ways to do this (PCA, weighted PCA...) and they will be presented in the next section. Note that we assume that the dimension p of E is known, in practice it can be estimated using hypothesis testing [19].

As the ADE method [13], the method we have just described is based on the integration by part formula (10). As a result, our method may be seen as a new version of ADE, called average derivative estimator by test functions (ADETF). The main difference between ADE and ADETF is that ADE puts $\psi = f$ so that their estimates only recover a single direction. This problem has been circumvented in the recent study [28] where the authors consider $\psi = \widetilde{\psi}\nabla f + \nabla\widetilde{\psi} f$ for some $\widetilde{\psi}$. First, by considering different functions ψ, our estimator is able to recover the multiple index. Secondly, comparing to both latter references, our approach does not need to estimate the derivatives of the density, and as a result does not require to select two different bandwidths. Moreover the presence of $\nabla\widehat{f}$ in ADE may induce an unnecessary noise that could affect badly the estimation. In the asymptotic variance of ADE

$$\text{var}\left(\nabla g(X) + \frac{(Y - g(X))\nabla f(X)}{f(X)}\right), \tag{12}$$

see Theorem 3.1 of [13], this is reflected by the additional term $\nabla g(X)$ that does not affect the variance of ADETF provided in Corollary 3.

4.2 Parameter setting

Choice of the bandwidth and the kernel. Theoretical results provided by Corollary 3 require the use of a high order kernel to reduce the bias. Since our simulations have highlighted that the use of high order kernels are not as crucial in practice as in theory, we consider the Epanechnikov radial kernel given by

$$K(x) \propto (1 - \|x\|^2),$$

such that $\int K = 1$. Contrarily to ADE, it turns out that ADETF is not really affected by the choice of the bandwidth. As a result, in the whole study, we select the optimal bandwidth for ADE and we put $h = 2sn^{-1/(d+2)}$ for ADETF, where s is the estimated standard deviation of X.

Choice of the test functions. We define

$$\psi(x) = \widetilde{\psi}(h_0^{-1}\|x\|) \qquad \text{with} \quad \widetilde{\psi}(z) = (1-z)^2(1+z)^2 1_{\{|z|<1\}}$$

where the scaling parameter h_0 is equal to the empirical estimator of $s = \mathbb{E}[\|X - \mathbb{E}[X]\|^2]^{1/2}$, and our test functions are

$$\psi_k(x) = \psi(x - t_k), \quad k = 1, \dots K.$$

Observing that better results are obtained if we do not restrict ourselves to a small value of K, we ended up with the simple choice $t_k = X_k$.

Computation of the directions. We have to extract p directions from $(\widehat{\beta}_k)_{k=1,\dots,n}$. Two approaches can be used and combined.

a) Use a criterion of dependence between Y and $\widehat{\beta}_k^T X$ to select among the $\widehat{\beta}_k$'s.

b) Choose the best direction through a PCA of these vectors.

The set $(\widehat{\beta}_k)_{k=1,\dots,n}$ is an heterogeneous family of estimated vector. Indeed because our choice was to visit every design point with the functions ψ_k's (in order not to loose information), some vectors in $(\widehat{\beta}_k)_{k=1,\dots,n}$ have a high variance and a large bias. To cancel their bad effect, we conduct step a) by selecting the root n vectors among the β_k's that have the larger dependence criterion

$$\sum_{h,h'} \frac{\left(p_{hh'} - \overline{p_{hh'}}^h \ \overline{p_{hh'}}^{h'}\right)^2}{\overline{p_{hh'}}^h \ \overline{p_{hh'}}^{h'}}$$

where $p_{h,h'} = \frac{1}{n}\sum_{i=1}^n 1_{\{Y_i \in I_h\}} 1_{\{(\beta_k^T X_i) \in J_{h'}\}}$ and $\overline{\cdot}^h$ is the mean over h. The partitions (I_h) and (J_h) have been defined having $\ulcorner\sqrt{(n)}\urcorner$ elements with equal sized (except the last). After this refinement we conduct step b), i.e. a PCA on the remaining vector $(\beta_k)_{k\in S}$. That is our final estimate of the *index* is given by the p eigenvectors of

$$\sum_{k\in S} \beta_k \beta_k^T$$

associated with the p-largest eigenvalues.

4.3 Simulations

The ADETF method follows a typical semiparametric approach characterized by mild assumptions on the design but that requires the nonparametric estimation of the density. In a different spirit, a well known competitor is the approach called *inverse regression* [17], that needs the *linearity condition* (slightly weaker than ellipticity of the distribution of X_1). In the following simulation study, we compare the estimation of the index space E given by ADE and ADETF with the one given by inverse regression methods, namely *Sliced inverse regression* (SIR) [17] and *Sliced average variance estimation* (SAVE) [3]. One remarks that in the whole simulation study, the predictors are drawn from the Gaussian distribution. This is quite a comfortable situation for SIR and SAVE since they are not penalized by the restrictive framework they impose. For each estimate $\widehat{E}$ of E, we compute the estimation error with

$$\|\widehat{P} - P\|_{\mathrm{F}}, \tag{13}$$

where P (resp. $\widehat{P}$) is the orthogonal projector on E (resp. $\widehat{E}$) and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. In each situation, we assume that the dimension of E is known.

4.3.1 The models

Model I. We first consider

$$Y = (\beta^T X)\sin\left(\beta^T X\right) + e,$$

where $X = (X^{(1)}, \ldots, X^{(p)}) \overset{\mathrm{d}}{=} \mathcal{N}(0, I)$, $e \overset{\mathrm{d}}{=} \mathcal{N}(0,1)$. It is well known [3] that the SIR method fails when the link function is symmetric whereas SAVE achieves consistency. As a result, we run ADE, ADETF and SAVE on Model I with different values of the parameters n and p. The boxplot are provided in Figure 2.

Model II. From now we fix $p = 6$ and $n = 200$ (this illustrates situations quite difficult) and we focus on different link functions, each representing interested situations. In order to better understand how do the symmetries in the link function influence the methods, we generate

$$Y = \cos\left(\frac{\pi}{2}(X^{(1)} - \mu)\right) + 0.5e,$$

with $\mu \in \mathbb{R}$. In our simulation, we try different values of μ from 0, which correspond to a symmetric link function, to 1. The boxplots are provided in Figure 3.

Model III. To highlight how the methods behave facing link functions with different level of fluctuations, we consider

$$\text{Model III:} \qquad Y = \tau \sin\left(X^{(1)}/\tau\right) + 0.5e,$$

with $\tau \in \mathbb{R}$. For different values of τ, we provide the boxplots of the errors in Figure 4.

Model IV. We conclude by a two dimensional model defined as

$$\text{Model IV:} \qquad Y = \frac{\sin(2X^{(1)})}{.5 + |1 + X^{(2)}|} + \sigma e,$$

where we found interesting to consider different values of σ. The method ADE does not appear because it only estimates a single direction.

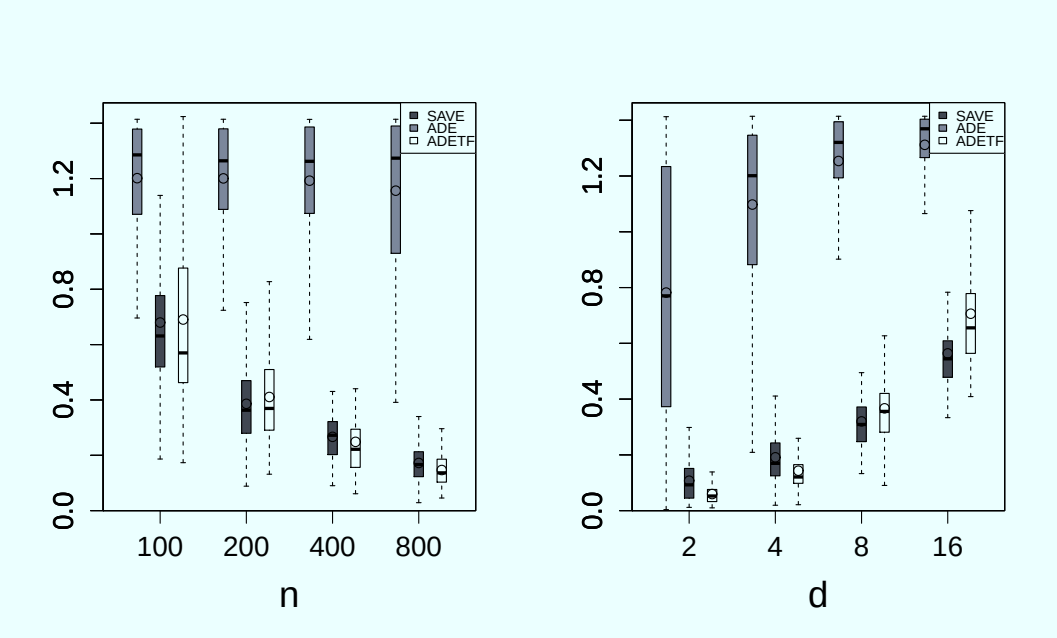

Figure 2: Boxplot over 100 samples of the estimation error (13) of SAVE, ADE and ADETF in the case of Model I, for different values of d (when $n = 400$) and different values of n (when $d = 6$).

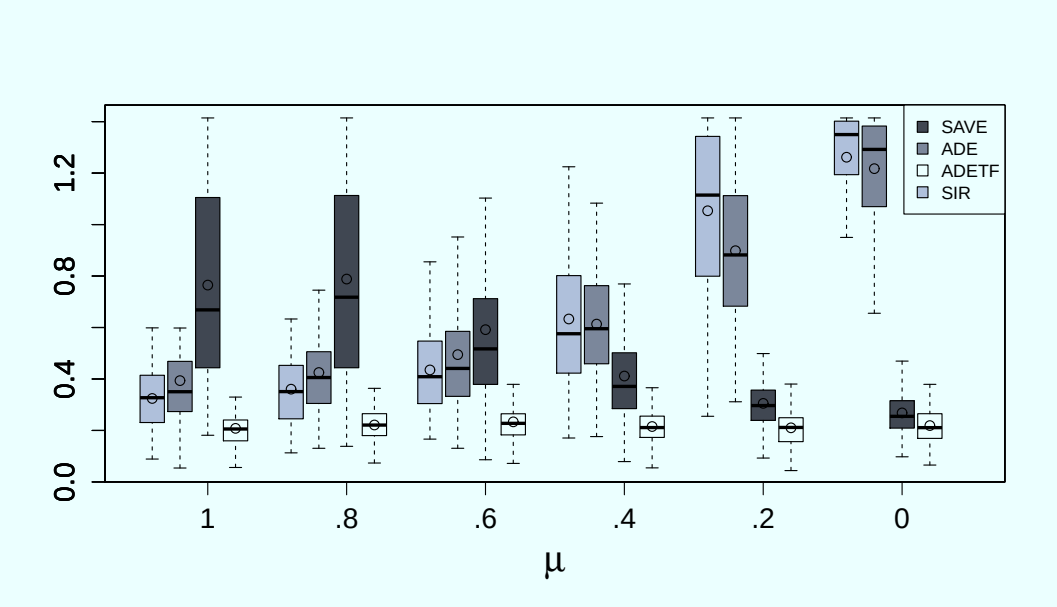

Figure 3: Boxplot over 100 samples of the estimation error (13) of SIR, SAVE, ADE and ADETF in the case of Model II, when $n = 200$ and for different values of μ.

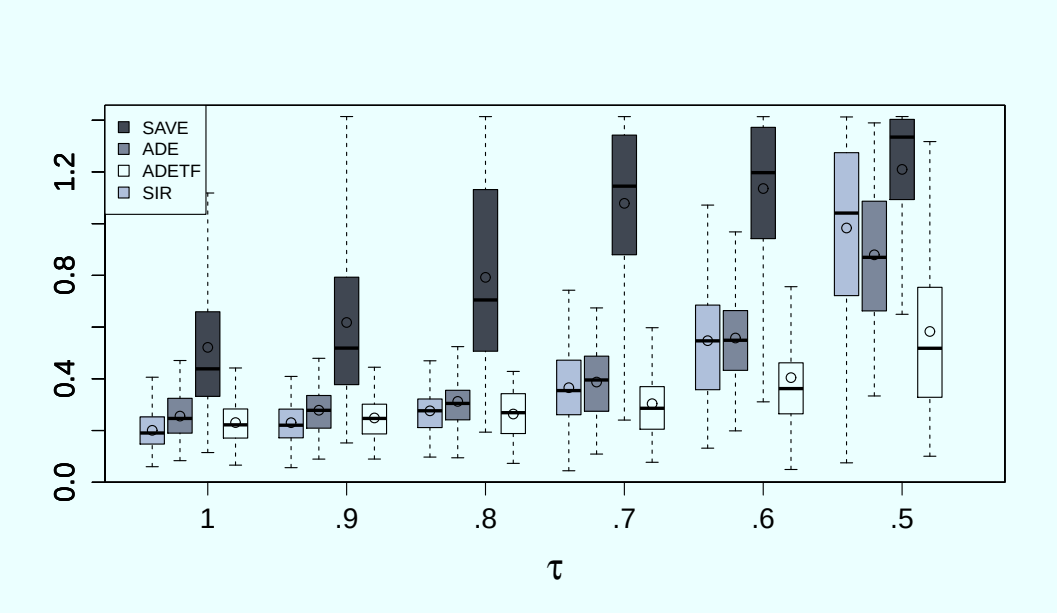

Figure 4: Boxplot over 100 samples of the estimation error (13) of SIR, SAVE, ADE and ADETF in the case of Model III, when $n = 200$ and for different values of τ.

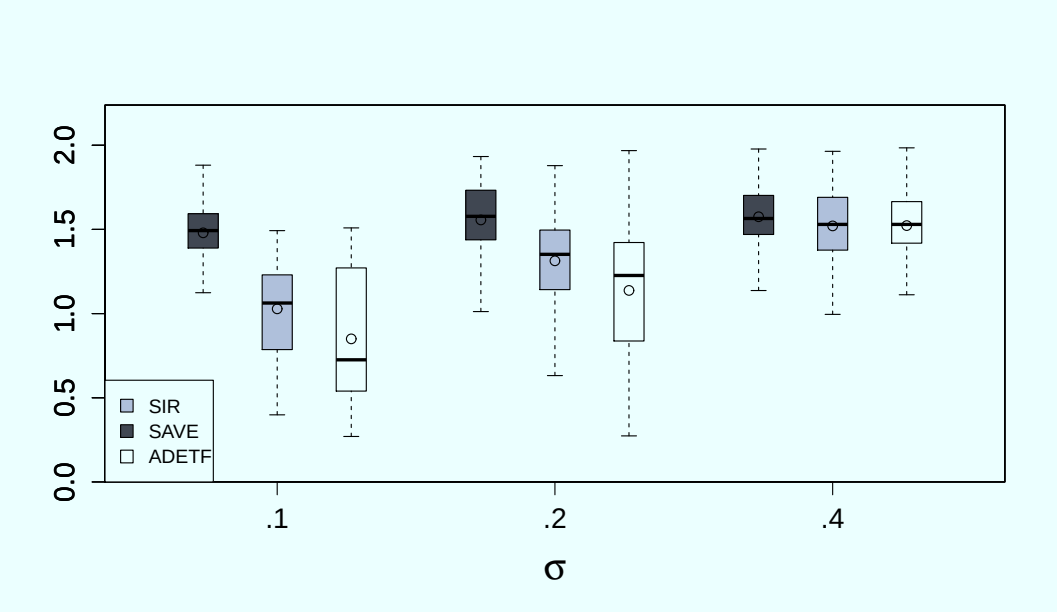

Figure 5: Boxplot over 100 samples of the estimation error (13) of SIR, SAVE, ADE and ADETF in the case of Model IV, when $n = 200$ and for different values of σ.

4.3.2 Interpretation of the results

In figure 2, one remarks the accuracy of SAVE and ADETF whereas ADE fails completely to estimate the index. Asymptotically, ADETF becomes better than SAVE whereas SAVE seems to be more robust than ADETF when d increase. The reason for this behavior when d increases is the so called *curse of dimensionality* raised in Remark 1.

In figure 3, we analyse more in details how does the symmetry impact the methods. We remark that SIR and ADE produce similar poor estimate when the link function is symmetric. On the other hand, while SAVE is consistent in the presence of symmetry it seems to fail when the function is odd. Indeed, whereas SAVE and SIR and ADE seem to perform symmetrically with respect to the value of μ, ADETF remains stable.

In figure 4 and 5, we see that ADETF is more robust to the variation of the scale than other methods as SIR or ADE. In the two dimensional model, one may see that ADETF produce the better estimate for every level of noise considered.

4.4 Adaptive ADE

Unfortunately ADE and ADETF are subject to the so called *curse of dimensionality*. As highlighted in Remark 1, the larger the dimension d the smoother the density f needs to be. Moreover, even if the density is smooth enough, one needs to use a high order kernel that may has poor performance at small sample size. In order to minimize bad effects of high dimension, we introduce the following adaptive strategy.

In [14] the authors proposed to estimate β by an averaging of ∇g using a *local linear estimator* [7] of g. In order to attain the root n consistency, their estimator needs to be improved via an adaptive procedure. The idea is simple: once β is estimated, one could think of running once more the estimation procedure in the reduced space in order to get advantage of the dimension reduction. The point is this cannot be done exactly since the reduction space remains unknown; however the authors proved that using an estimate of β with a suitable implementation, this idea is fruitful theoretically as well as practically.

All this is in theory not necessary in our case since, if f is regular enough, the root n consistency is achieved whatever the dimension, but we observe that this refinement procedure gives good results in practice. Following their idea we notice that for any test function ψ

$$\mathbb{E}\left[\frac{Y_1 A\nabla\psi(AX_1)}{f_{|AX_1}(AX_1)}\right] = -\mathbb{E}\left[\frac{\nabla g(X_1)\psi(AX_1)}{f_{|AX_1}(AX_1)}\right] \in E \qquad \text{provided that } E \subset \text{span}(A), \tag{14}$$

where $f_{|AX_1}$ is the density of AX_1. For any A we have the estimator

$$\widehat{\beta}_\psi(A) = n^{-1}\sum_{i=1}^{n} \frac{Y_i A\nabla\psi(AX_i))}{\widehat{f}_{|AX_1}(AX_i)}, \tag{15}$$

with

$$\widehat{f}_{|AX_1}(x) = (nh^d)^{-1}\sum_{i=1}^{n} K(h^{-1}(AX_i - x)), \qquad \text{for every } x \in \mathbb{R}^p.$$

After an initial estimation $\widehat{\beta}$ obtained with $A = Id$ and several test functions $\psi_1, \dots \psi_K$, we take $A = \widehat{\beta}\widehat{\beta}^T + \epsilon I$ as in [14] and obtain a second estimator whose window has been stretched in the interesting direction, i.e. the direction where g varies. This procedure might be iterated several times with h and ϵ decreasing.

The theoretical study and the implementation details require much more work that seems to be beyond the scope of the present paper. This could be done following the well documented semiparametric literature on the subject [14], [4] and [27].

5 A remark about the generalization of Theorem 1

In view of the intriguing convergence rates stated in Theorem 1, one may be curious to know the behavior of our estimator when estimating more general functionals with the form

$$I_T = \int T(x, f(x))dx,$$

where $T : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}$ is such that $y \mapsto T(x, y)$ has a second order derivative bounded uniformly on x. Following the approach of Section 2, the estimator we consider is

$$\widehat{I}_T = n^{-1} \sum_{i=1}^{n} \frac{T\big(X_i, \widehat{f}^{(i)}(X_i)\big)}{\widehat{f}^{(i)}(X_i)}. \tag{16}$$

The study of the asymptotic behavior of $\sqrt{n}(\widehat{I}_T - I_T)$ generalizes Theorem 1. It turns out that the case $T : (x, y) \mapsto \varphi(x)$ is the only case where the rates are faster than root n. For other functionals, $\sqrt{n}(\widehat{I}_T - I_T)$ converges to a normal distribution. In view of the negative aspect of the following results with respect to those of Theorem 1, we provide an informal calculation that leads to the asymptotic law of $\sqrt{n}(\widehat{I}_T - I_T)$. By assumption on T, using a Taylor expansion with respect to the second coordinate of T, we have

$$n^{1/2}(\widehat{I}_T - I_T) = n^{-1/2} \sum_{i=1}^{n} \left(\frac{T(X_i, f(X_i))}{\widehat{f}_i} - I_T + \frac{\partial_y T(X_i, f(X_i))(\widehat{f}_i - f(X_i))}{\widehat{f}_i} \right) + \widehat{R}_2,$$

with

$$|\widehat{R}_2| \leq C n^{-1/2} \sum_{i=1}^{n} \frac{(\widehat{f}_i - f(X_i))^2}{\widehat{f}_i} = O_{\mathbb{P}}(n^{1/2} h^{2r} + n^{-1/2} h^{-d})$$

due to equations (22) and (28). Then, we write

$$\sqrt{n}(\widehat{I}_T - I_T) = \widehat{R}_0 + \widehat{R}_1 + \widehat{R}_2,$$

with

$$\widehat{R}_0 = n^{-1/2} \sum_{i=1}^{n} \frac{T(X_i, f(X_i))}{\widehat{f}_i} - I_T - \frac{\partial_y T(X_i, f(X_i)) f(X_i)}{\widehat{f}_i} + \int \partial_y T(x, f(x)) f(x) dx$$

$$\widehat{R}_1 = n^{-1/2} \sum_{i=1}^{n} \partial_y T(X_i, f(X_i)) - \int \partial_y T(x, f(x)) f(x) dx.$$

Provided Theorem 1 can be applied two times, we show that $\widehat{R}_0 = o_{\mathbb{P}}(1)$. As a consequence $\sqrt{n}(\widehat{I}_T - I_T) = o_{\mathbb{P}}(1)$ if and only if the variance of $\widehat{R}_1$ is degenerate, that is equivalent to

$$\partial_y T(X_i, f(X_i)) = c \qquad \text{a.s.}$$

If we want this to be true for a reasonably large class of distribution function, it would imply

$$\partial_y T(x, y) = c \qquad \text{for all } (x, y) \in \mathbb{R}^d \times \mathbb{R}^+,$$

the solutions have the form $T(x, y) = \varphi(x) + cy$.

6 Concluding remarks

There exists some links between Theorem 2 and nonparametric estimation. Those links are beyond the scope of this article but can be the subject of further research. Indeed, Theorem 2 is not so far from dealing with nonparametric regression. On the one hand, one can use it for the estimation of the Fourier (or wavelet) coefficient in the L_2 expansion of g

$$\widehat{c}_k(g) = n^{-1} \sum_{i=1}^{n} \frac{Y_i \psi_k(X_i)}{\widehat{f}^{(i)}(X_i)},$$

(ψ_k is the Fourier L_2 basis), this would lead to projection estimates

$$\sum_{k=1}^{K} \widehat{c}_k(g) \psi_k(y), \tag{17}$$

and estimates by shrinkage. On the other hand, one can similarly define the kernel estimator

$$\sum_{i=1}^{n} \frac{Y_i K_{h_2}(X_i - x)}{\sum_{j=1}^{n} K_{h_1}(X_j - X_i)}, \tag{18}$$

where h_1 and h_2 are bandwidths each linked with the estimation of f and the regularization of g, respectively. Similar estimators of the regression function have already been introduced in the case of unknown random design (density f). Estimate (17) is linked with the estimate (3.3.6) p.51 of [12], studied in [22], whereas estimate (18) is reminiscent of the Gasser-Muller estimator [10]. Both latter estimates are called convolution estimator of the regression because they estimate directly $< g, K_h(\cdot - y) >$ whereas the most popular approach, inspired by the Naradaya-Watson estimate, has been to estimate separately $< gf, K_h(\cdot - y) >$ and $< f, K_h(\cdot - y) >$ by simple empirical means, $\widehat{gf}$ and $\widehat{f}$ respectively, and then to estimate g by $\widehat{gf}/\widehat{f}$. It would be interesting to understand how Equation (17) or (18) could improve the estimation of g, work along this line is under progress.

7 Proofs

7.1 Proof of Theorem 1

For clarity, we introduce the following notation

$$\begin{aligned}
K_{ij} &= h^{-p} K(h^{-1}(X_i - X_j)) \\
\widehat{f}_i &= \frac{1}{n-1} \sum_{j \neq i}^{n} K_{ij} \\
\widehat{v}_i &= \frac{1}{(n-1)(n-2)} \sum_{j \neq i}^{n} (K_{ij} - \widehat{f}_i)^2,
\end{aligned}$$

and for any function $g : \mathbb{R}^p \to \mathbb{R}$, we define

$$g_h(x) = \int g(x + hu) K(u) du. \tag{19}$$

We start by showing (6), then (5) will follow straightforwardly.
Proof of (6):

The following development reminiscent of the Taylor expansion

$$\frac{1}{\widehat{f}_i} = \frac{1}{f_h(X_i)} + \frac{f_h(X_i) - \widehat{f}_i}{f_h(X_i)^2} + \frac{(f_h(X_i) - \widehat{f}_i)^2}{f_h(X_i)^3} + \frac{(f_h(X_i) - \widehat{f}_i)^3}{\widehat{f}_i f_h(X_i)^3},$$

allows to expand our estimator as a sum of many terms where the density estimate $\widehat{f}_i$ is moved to the numerator, with the exception of the fifth one. We will show that this last term goes quickly to 0. For the linearised terms, this is very messy because the correct bound will be obtained by expanding also $\widehat{f}_i$ in those expressions. In order to sort out these terms, we borrow to Vial [26] the trick of making appear a degenerate U-statistic in such a development (by inserting the right quantity in $\widehat{R}_0$ below). More explicitly, recalling that

$$n^{1/2}\left(\widehat{I}_{cor}(\varphi) - \int \varphi(x)dx\right) = n^{-1/2}\left(\sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}_i}\Big(1 - \frac{\widehat{v}_i}{\widehat{f}_i^2}\Big) - \int \varphi(x)dx\right),$$

using the notations

$$\psi_q(x) = \frac{\varphi(x)}{f_h(x)^q}, \quad q \in \mathbb{N},$$
$$\tilde{\psi}_1(x) = \left(\varphi(x)\frac{f(x)}{f_h(x)^2}\right)_h,$$

we obtain

$$n^{1/2}\left(\widehat{I}_{cor}(\varphi) - \int \varphi(x)dx\right) = \widehat{R}_0 + \widehat{R}_1 + \widehat{R}_2 + \widehat{R}_3 + \widehat{R}_4 + \widehat{R}_5 \tag{20}$$

with (we underbrace terms which have been deliberately introduced and removed)

$$\widehat{R}_0 = n^{-1/2}\sum_{i=1}^{n} \psi_1(X_i) - \psi_2(X_i)\widehat{f}_i + \underbrace{\tilde{\psi}_1(X_i)} - \underbrace{\mathbb{E}[\psi_1(X_i)]}$$
$$\widehat{R}_1 = \int \Big(\underbrace{f(x)f_h(x)^{-1}} - 1\Big)\varphi(x)\,dx$$
$$\widehat{R}_2 = n^{-1/2}\sum_{i=1}^{n} \psi_1(X_i) - \underbrace{\tilde{\psi}_1(X_i)}$$
$$\widehat{R}_3 = n^{-1/2}\sum_{i=1}^{n} \psi_3(X_i)\{(f_h(X_i) - \widehat{f}_i)^2 - \underbrace{\widehat{v}_i}\}$$
$$\widehat{R}_4 = n^{-1/2}\sum_{i=1}^{n} \frac{\psi_3(X_i)\widehat{v}_i}{\widehat{f}_i^3}(\underbrace{\widehat{f}_i^3} - f_h(X_i)^3)$$
$$\widehat{R}_5 = n^{-1/2}\sum_{i=1}^{n} \psi_3(X_i)\frac{(f_h(X_i) - \widehat{f}_i)^3}{\widehat{f}_i}.$$

$\widehat{v}_i$ appears to be a centering term in $\widehat{R}_3$. We shall now compute bounds for each term separately. Since some of these bound will be used for the proof of (5) we shall use only the property

$$h^s + n^{1/2}h^r + n^{-1/2}h^{-d} \to 0.$$

Step 1 : $\|\widehat{R}_0\|_2 = O(n^{-1/2}h^{-d/2})$. Remark that

$$\widehat{R}_1 = n^{-1/2}(n-1)^{-1}\sum_{i\neq j}\mathbb{E}[u_{ij}|X_j] - u_{ij} + E[u_{ij}|X_i] - E[u_{ij}],$$

with $u_{ij} = \psi_2(X_i)K_{ij}$, is a degenerate $U-$statistic. The $n(n-1)$ terms in the sum are all orthogonal with L_2 norm smaller than $\|u_{ij}\|_2$, hence

$$(n-1)E[\widehat{R}_1^2] \leq \mathbb{E}[u_{12}^2] \leq \|\psi_2\|_\infty^2 E[K_{12}^2]$$

and

$$\mathbb{E}[K_{12}^2|X_1] \leq h^{-2d}\int K(h^{-1}(x-X_1))^2 f(x)dx \leq h^{-d}\|f\|_\infty \int K(u)^2 du. \tag{21}$$

Step 2 : $\widehat{R}_1 = O(n^{1/2}h^r)$. This classically results from Equation (29) of Lemma 4, and from Assumption (B3).

Step 3 : $\|\widehat{R}_2\|_2 = O(n^{1/2}h^r + h^s)$. We can rearrange the function $\psi_1(x) - \tilde{\psi}_1$ as

$$\psi_1(x) - \tilde{\psi}_1(x) = \Big(\psi_1(x) - \psi_{1h}(x)\Big) + \Big(\psi_{1h}(x) - \tilde{\psi}_1(x)\Big)$$

(with the notation (19)) and since

$$\begin{aligned}\|\psi_{1h}(x) - \tilde{\psi}_1(x)\|_\infty =& \|\Big(\psi_1(x) - \varphi(x)\frac{f(x)}{f_h(x)^2}\Big)_n\|_\infty \\ \leq& \|\psi_1(x) - \varphi(x)\frac{f(x)}{f_h(x)^2}\|_\infty \\ =& \Big\|\frac{\varphi}{f_h^2}(f_h - f)\Big\|_\infty\end{aligned}$$

we have

$$\widehat{R}_2 \leq n^{-1/2}\Big|\sum_{i=1}^n \psi_{1h}(X_i) - \psi_1(X_i)\Big| + n^{1/2}\Big\|\frac{\varphi}{f_h^2}\Big\|_\infty \|f_h - f\|_\infty$$

and we conclude with Equations (30) and (29) of Lemma 4.

Step 4 : $\|\widehat{R}_3\|_2 = O(n^{-1/2}h^{-d/2})$. We first rewrite separately each term. Set

$$U_i = (f_h(X_i) - \widehat{f}_i)^2 - \widehat{v}_i,$$

and rewrite $\widehat{R}_3$ as

$$\widehat{R}_3 = n^{-1/2}\sum_{i=1}^n \psi_3(X_i)U_i.$$

Consider a sequence of real numbers $(x_j)_{1\leq j\leq p}$ and set

$$\begin{aligned}m &= \frac{1}{p}\sum_{j=1}^p x_j \\ v &= \frac{1}{p(p-1)}\sum_{j=1}^p (x_j - m)^2 = \frac{1}{p(p-1)}\sum_{j=1}^p (x_j^2 - m^2),\end{aligned}$$

then

$$m^2 - v = \Big(1 + \frac{1}{p-1}\Big)m^2 - \frac{1}{p(p-1)}\sum_{j=1}^{p} x_j^2 = \frac{2}{p(p-1)}\sum_{j<k} x_j x_k.$$

Applying this with $x_j = K_{ij} - f_h(X_i)$ (i is fixed) and $p = n-1$ we get

$$\begin{aligned} U_i =& \frac{2}{(n-1)(n-2)} \sum_{j\neq i, k\neq i, j<k} (K_{ij} - f_h(X_i))(K_{ik} - f_h(X_i)) \\ =& \frac{2}{(n-1)(n-2)} \sum_{j<k} \xi_{ij}\xi_{ik}, \end{aligned}$$

with

$$\begin{aligned} \xi_{ij} &= K_{ij} - f_h(X_i) \\ \xi_{ii} &= 0. \end{aligned}$$

Then

$$\widehat{R}_3 = \frac{2n^{-1/2}}{(n-1)(n-2)} \sum_i \sum_{j<k} \psi_3(X_i)\xi_{ij}\xi_{ik}.$$

We are going to calculate $\mathbb{E}[\widehat{R}_3^2]$ by using the Efron-Stein inequality (Theorem 5) and the moment inequalities (33) to (35) for ξ_{ij} stated in Lemma 7; in particular, by (33) $\mathbb{E}[\widehat{R}_3^2] = Var(\widehat{R}_3)$. Consider $\widehat{R}_3 = f(X_1, \dots X_n)$ as a function of the X_i's and define

$$\begin{aligned} \widehat{R}_3' &= f(X_1', X_2 \dots X_n) \\ \xi_{1j}' &= h^{-d}K(h^{-1}(X_1' - X_i)) - f_h(X_1) \\ \xi_{i1}' &= h^{-d}K(h^{-1}(X_1' - X_i)) - f_h(X_i) \\ \xi_{ij}' &= \xi_{ij} \text{ if } i \neq 1 \text{ and } j \neq 1 \end{aligned}$$

where X_1' is a copy of X_1 independent from the sample $(X_1, \dots, X_n)$. Then by the Efron-Stein inequality and the triangular inequality

$$\begin{aligned} \|\widehat{R}_3\|_2 \leq& \Big(\frac{n}{2}\Big)^{1/2} \|\widehat{R}_3 - \widehat{R}_3'\|_2 \\ =& Cn^{-2} \left\| \sum_{j<k} (\psi_3(X_1)\xi_{1j}\xi_{1k} - \psi_3(X_1')\xi_{1j}'\xi_{1k}') + \sum_i \sum_{1<k} \psi_3(X_i)(\xi_{i1} - \xi_{i1}')\xi_{ik} \right\| \\ \leq& Cn^{-2} \left(\left\| \sum_{j<k} \psi_3(X_1)\xi_{1j}\xi_{1k} - \psi_3(X_1')\xi_{1j}'\xi_{1k}' \right\| + \left\| \sum_{1<k} \sum_i \psi_3(X_i)(\xi_{i1} - \xi_{i1}')\xi_{ik} \right\| \right) \\ =& Cn^{-2}(\|T_1\|_2 + \|T_2\|_2). \end{aligned}$$

Remember that $\xi_{ii} = 0$. Noting that the terms in the first sum are orthogonal (by independence of ξ_{ij} and ξ_{ik} conditionally to X_i and (33)) we obtain

$$\begin{aligned} \|T_1\|_2 =& \sqrt{\tfrac{(n-1)(n-2)}{2}} \|\psi_3(X_1)\xi_{12}\xi_{13} - \psi_3(X_1')\xi_{12}'\xi_{13}'\|_2 \\ \leq& \sqrt{2}\, n \|\psi_3\|_\infty \|\xi_{12}\xi_{13}\|_2 \\ =& \sqrt{2}\, n \|\psi_3\|_\infty \mathbb{E}[\mathbb{E}[\xi_{12}^2\xi_{13}^2|X_1]]^{1/2} \\ =& \sqrt{2}\, n \|\psi_3\|_\infty \|\mathbb{E}[\xi_{12}^2|X_1]\|_2 \\ \leq& Cnh^{-d} \end{aligned}$$

by (34). Because the terms of the second sum are orthogonal whenever the values of k are different, we get

$$\|T_2\|_2 = (n-1)^{1/2} \Big\| \sum_i \psi_3(X_i)(\xi_{i1} - \xi'_{i1})\xi_{i2} \Big\|.$$

By first developing and then using that X'_1 is an independent copy of X_1, we obtain

$$\begin{aligned}
\Big\| \sum_i \psi_3(X_i)(\xi_{i1} - \xi'_{i1})\xi_{i2} \Big\|_2^2 \leq & n\mathbb{E}\left[\psi_3(X_3)^2(\xi_{31} - \xi'_{31})^2\xi_{32}^2\right] | \\
& + n^2 |\mathbb{E}\left[\psi_3(X_3)\psi_3(X_4)(\xi_{31} - \xi'_{31})\xi_{32}(\xi_{41} - \xi'_{41})\xi_{42}\right] | \\
\leq & nC\mathbb{E}\left[(\xi_{31} - \xi'_{31})^2\xi_{32}^2\right] \\
& + n^2 C'\mathbb{E}\left[|\mathbb{E}[(\xi_{31} - \xi'_{31})\xi_{32}(\xi_{41} - \xi'_{41})\xi_{42}|X_3, X_4]|\right] \\
= & 2Cn\mathbb{E}\left[\xi_{31}^2\xi_{32}^2\right] + 2Cn^2\mathbb{E}\left[|\mathbb{E}[\xi_{31}\xi_{32}\xi_{41}\xi_{42}|X_3, X_4]|\right].
\end{aligned}$$

Then by (34) $\mathbb{E}\left[\xi_{31}^2\xi_{32}^2\right] = \mathbb{E}\left[\mathbb{E}[\xi_{31}^2|X_3]^2\right] \leq Ch^{-2d}$ and by (35)

$$\begin{aligned}
\mathbb{E}[|\mathbb{E}[\xi_{31}\xi_{32}\xi_{41}\xi_{42}|X_3, X_4]|] = & \mathbb{E}[\mathbb{E}[\xi_{31}\xi_{41}|X_3, X_4]^2] \\
\leq & 2\|f\|_\infty^2 h^{-2d}\mathbb{E}[K_2(h^{-1}(X_4 - X_3))^2] + 2\|f\|_\infty^4 \\
\leq & 2\|f\|_\infty^3 h^{-d} \int K_2(u)^2 du + 2\|f\|_\infty^4.
\end{aligned}$$

Bringing everything together

$$\|\widehat{R}_3\|_2 \leq Cn^{-1}h^{-d} + Cn^{-1}h^{-d} + Cn^{-1/2}h^{-d/2} = O(n^{-1/2}h^{-d/2})$$

because $nh^d \to \infty$.

Step 5 : $\widehat{R}_4 = O_{\mathbb{P}}(n^{-1}h^{-3d/2})$. We start with a lower bound for $\widehat{f}_i$ by proving the existence of $N(\omega)$ such that

$$\forall\, n \geq N(\omega),\ \forall\, i,\quad \frac{b}{2} < \widehat{f}_i < 2\|f\|_\infty. \tag{22}$$

Notice that

$$\begin{aligned}
\widehat{f}_i &= \frac{n}{n-1}\left(\widehat{f}(X_i) - \frac{h^{-d}}{n-1}K(0)\right) \\
\widehat{f}(x) &= \frac{1}{nh^d}\sum_{k=1}^n K(h^{-d}(x - X_k)),
\end{aligned}$$

due to the almost sure uniform convergence of $\widehat{f}$ to f (Theorem 1 in [6]) we have for n large enough

$$\frac{2b}{3} < \inf_{x\in Q} \widehat{f}(x) \leq \sup_{x\in Q} \widehat{f}(x) < \frac{3}{2}\|f\|_\infty$$

and since assumption $nh^d \to \infty$, (22) follows. We can now compute the expectation of $\widehat{R}_4$ restricted to $\{n \geq N(\omega)\}$. Because

$$|\widehat{R}_4|1_{n>N(\omega)} \leq Cn^{-1/2}\sum_{i=1}^n |\widehat{f}_i - f_h(X_i)|\widehat{v}_i$$

we have by the Cauchy-Schwartz inequality

$$\mathbb{E}[|\widehat{R}_4|1_{n>N(\omega)}] \leq Cn^{1/2}\mathbb{E}[(\widehat{f}_1 - f_h(X_1))^2]^{1/2}\mathbb{E}[\widehat{v}_1^2]^{1/2}. \tag{23}$$

Applying the fact that for any real number a, $\frac{1}{p}\sum_{j=1}^{p}(x_j - \overline{x})^2 \leq \frac{1}{p}\sum_{i=1}^{p}(x_j - a)^2$ to $x_j = K_{1j}$, $p = n-1$ and $a = f_h(X_1)$, we obtain that

$$\widehat{v}_1 \leq \frac{1}{(n-1)(n-2)}\sum_{j=2}^{n}\xi_{1j}^2,$$

then using (34)

$$\begin{aligned}\mathbb{E}[\widehat{v}_1^2] =& (n-1)^{-1}(n-2)^{-2}\mathbb{E}[\xi_{12}^4] + (n-1)^{-1}(n-2)^{-1}\mathbb{E}[\xi_{12}^2\xi_{13}^2] \\ \leq& C'n^{-3}h^{-3d} + C'n^{-2}h^{-2d} \\ \leq& C''n^{-2}h^{-2d}\end{aligned} \tag{24}$$

because nh^d is lower bounded. On the other hand using (34)

$$\mathbb{E}[(\widehat{f}_1 - f_h(X_1))^2] = \frac{1}{n-1}\mathbb{E}[\xi_{1i}^2] \leq Cn^{-1}h^{-d}. \tag{25}$$

Putting together (23), (24) and (25),

$$\mathbb{E}[|\widehat{R}_4|1_{n>N(\omega)}] \leq Cn^{1/2}n^{-1}h^{-d}n^{-1/2}h^{-d/2} = Cn^{-1}h^{-3d/2}.$$

In particular

$$\begin{aligned}\mathbb{P}(nh^{3d/2}|\widehat{R}_4| > A) \leq& \mathbb{P}(nh^{3d/2}|\widehat{R}_4|1_{n>N(\omega)} > A) + \mathbb{P}(n \leq N(\omega)) \\ \leq& C\,A^{-1} + \mathbb{P}(n \leq N(\omega)).\end{aligned}$$

This proves the boundedness in probability of $nh^{3d/2}|\widehat{R}_4|$.

Step 6 : $\widehat{R}_5 = O_{\mathbb{P}}(n^{-1}h^{-3d/2} + n^{-3/2}h^{-2d})$. Following (22) since

$$|\widehat{R}_5|1_{n>N(\omega)} \leq 2b^{-3}\|\varphi\|_\infty n^{-1/2}\sum_{i=1}^{n}|\widehat{f}_i - f_h(X_i)|^3,$$

we can show the convergence in probability of the right-hand side term as in Step 5. We have indeed by the Rosenthal's inequality[1]

$$\begin{aligned}\mathbb{E}\left[n^{-1/2}\sum_{i=1}^{n}|\widehat{f}_i - f_h(X_i)|^p\right] =& n^{1/2}(n-1)^{-p}\mathbb{E}[|\sum_{i=2}^{n}\xi_{1i}|^p] \\ \leq& Cn^{1/2}n^{-p}\{(n\mathbb{E}[\xi_{12}^2])^{p/2} + n\mathbb{E}[|\xi_{12}|^p]\} \\ \leq& C'\{n^{(1-p)/2}h^{-pd/2} + n^{3/2-p}h^{-(p-1)d}\}.\end{aligned} \tag{26}$$

(cf. (34)). Hence with $p = 3$

$$\mathbb{E}\left[|\widehat{R}_5|1_{n>N(\omega)}\right] \leq C\{n^{-1}h^{-3d/2} + n^{-3/2}h^{-2d}\}$$

[1]For a martingale $(S_i, \mathcal{F}_i)_{i\in\mathbb{N}}$ and $2 \leq p < +\infty$, we have $\mathbb{E}[|S_n|^p] \leq C\{\mathbb{E}[(\sum_{i=1}^{n}\mathbb{E}[X_i^2|\mathcal{F}_{i-1}])^{p/2}] + \sum_{i=1}^{n}\mathbb{E}|X_i|^p\}$, where $X_i = S_i - S_{i-1}$ (see for instance [11], p. 23-24).

and we conclude as in Step 5.

Proof of (6): Putting together the steps 1 to 6, and taking into account, concerning $\widehat{R}_5$, that $n^{-3/2}h^{-2d} = (n^{-1/2}h^{-d/2})(n^{-1}h^{-3d/2})$, we obtain (6).

For (5), we use a shorter expansion which leads to an actually much simpler proof:

$$\frac{1}{\widehat{f}_i} = \frac{1}{f_h(X_i)} + \frac{f_h(X_i) - \widehat{f}_i}{f_h(X_i)^2} + \frac{(f_h(X_i) - \widehat{f}_i)^2}{\widehat{f}_i f_h(X_i)^2}$$

and

$$r = n^{-1/2}\Big(\sum_{i=1}^{n} \frac{\varphi(X_i)}{\widehat{f}_i} - \int \varphi(x)dx\Big) = \widehat{R}_0 + \widehat{R}_1 + \widehat{R}_2 + \widehat{R}'_5$$

with

$$\begin{aligned}
&\psi_q(x) = \frac{\varphi(x)}{f_h(x)^q}, \quad q \in \mathbb{N} \\
&\widehat{R}_0 = \int \Big(\underbrace{f(x)f_h(x)^{-1}}\,-1\Big)\varphi(x)\,dx \\
&\widehat{R}_1 = n^{-1/2}\sum_{i=1}^{n} \psi_1(X_i) - \psi_2(X_i)\widehat{f}_i + \underbrace{\tilde{\psi}_1(X_i)} - \underbrace{\mathbb{E}[\psi_1(X_i)]}, \quad \tilde{\psi}_1(x) = \left(\varphi(x)\frac{f(x)}{f_h(x)^2}\right)_h \\
&\widehat{R}_2 = n^{-1/2}\sum_{i=1}^{n} \psi_1(X_i) - \underbrace{\tilde{\psi}_1(X_i)} \\
&\widehat{R}'_5 = n^{-1/2}\sum_{i=1}^{n} \psi_2(X_i)\frac{(f_h(X_i) - \widehat{f}_i)^2}{\widehat{f}_i}.
\end{aligned}$$

The term $\widehat{R}'_5$ is bounded exactly as $\widehat{R}_5$ but since now we use (26) with $p = 2$ instead of $p = 3$, we obtain

$$\mathbb{E}[|\widehat{R}'_5|1_{n>N(\omega)}) \leq Cn^{1/2}\mathbb{E}[|f_h(X_1) - \widehat{f}_1|^2] \leq Cn^{-1/2}h^{-d}$$

and we get $|\widehat{R}'_5| = O_{\mathbb{P}}(n^{-1/2}h^{-d})$. □

7.2 Proof of the Theorem 2

By decomposition (7), we are interested in the asymptotic law of the vector

$$n^{-1/2}\sum_{i=1}^{n} \frac{\sigma(X_i)\psi(X_i)}{\widehat{f}_i}e_i + \; n^{-1/2}\left(\sum_{i=1}^{n} \frac{g(X_i)\psi(X_i)}{\widehat{f}_i} - \int g(x)\psi(x)dx\right).$$

By Lemma 1, the right hand-side term goes to 0 in probability. For the other term, we use the decomposition $\widehat{S}_1 + \widehat{S}_2$, with

$$\widehat{S}_1 = n^{-1/2}\sum_{i=1}^{n} \frac{s(X_i)}{f(X_i)}e_i \quad \text{and} \quad \widehat{S}_2 = n^{-1/2}\sum_{i=1}^{n} \frac{s(X_i)(f(X_i) - \widehat{f}(X_i))}{\widehat{f}_i f(X_i)}e_i. \tag{27}$$

where $s(X_i) = \sigma(X_i)\psi(X_i)$. We define $\mathcal{F}$ as the σ-field generated by the set of random variables $\{X_1, X_2, \cdots\}$. We get

$$\mathbb{E}[\widehat{S}_2^2|\mathcal{F}] = n^{-1}\sum_{i=1}^{n} \frac{s(X_i)^2(f(X_i) - \widehat{f}_i)^2}{\widehat{f}_i^2 f(X_i)^2},$$

then, one has

$$\mathbb{E}[\widehat{S}_2^2|\mathcal{F}] \leq (b^2 \inf_i \widehat{f}_i^2)^{-1} \|s\|_\infty^2 n^{-1} \sum_{i=1}^n (f(X_i) - \widehat{f}_i)^2.$$

For the term on the left, since s has support Q we can use (22), that is for n large enough, it is bounded. For the right hand-side term, it follows that

$$n^{-1} \sum_{i=1}^n (f(X_i) - \widehat{f}_i)^2 \leq 2(n^{-1} \sum_{i=1}^n (f(X_i) - f_h(X_i))^2 + n^{-1} \sum_{i=1}^n (f_h(X_i) - \widehat{f}_i)^2),$$

and then using Lemma 4 and (26) for $p = 2$ we provide the bound

$$\|n^{-1} \sum_{i=1}^n (f(X_i) - \widehat{f}_i)^2\|_1 \leq C\{h^{2r} + n^{-1}h^{-d}\}. \tag{28}$$

Therefore, we have shown that $\mathbb{E}[\widehat{S}_2^2|\mathcal{F}] \to 0$ in probability. Since for any $\epsilon > 0$, $\mathbb{P}(|\widehat{S}_2| > \epsilon|\mathcal{F}) \leq \epsilon^{-2}\mathbb{E}[\widehat{S}_2^2|\mathcal{F}]$, it remains to note that the sequence $\mathbb{P}(|\widehat{S}_2| > \epsilon|\mathcal{F})$ is uniformly integrable to apply the Lebesgue domination theorem to get

$$\mathbb{P}(\widehat{S}_2 > \epsilon) \longrightarrow 0.$$

To conclude, we apply the CLT to $\widehat{S}_1$ and the statement follows. □

7.3 Somme lemmas

Lemma 4. *For any function $g : \mathbb{R}^d \to \mathbb{R}$, we define*

$$g_h(x) = \int g(x + hu)K(u)du.$$

Under Assumptions (B1) and (B2) we have for some constant C

$$\|f_h - f\|_\infty \leq Ch^r, \tag{29}$$

and for any $\psi \in \mathcal{H}_s$, $\lfloor s \rfloor \leq r$ (cf Equation (4) and the following remark)

$$\left\| \sum_{i=1}^n \psi(X_i) - \psi_h(X_i) \right\|_2 \leq Cn^{1/2}(h^s + n^{1/2}h^r) \tag{30}$$

where C depends on ψ and f.

Proof. We split mean and variance:

$$\mathbb{E}[(\sum_{i=1}^n \psi(X_i) - \psi_h(X_i))^2] = (n\mathbb{E}[\psi(X_1) - \psi_h(X_1)])^2 + nVar(\psi(X_1) - \psi_h(X_1)).$$

For the mean:

$$\begin{aligned}
\mathbb{E}[\psi(X_1) - \psi_h(X_1)] &= \int (\psi(x) - \psi_h(x)) f(x)dx \\
&= \int \psi(x)f(x) - \psi(x)f_h(x)dx \\
&= \int \psi(x)(f(x) - f_h(x))dx \\
|\mathbb{E}[\psi(X_1) - \psi_h(X_1)]| &\leq Ch^r \|\psi\|_\infty
\end{aligned}$$

and for the variance

$$\mathbb{E}[(\psi_h(X_1) - \psi(X_1))^2] = \int \left(\int (\psi(x+hu) - \psi(x)) K(u) du \right)^2 f(x) dx. \tag{31}$$

By the Taylor formula with Lagrange remainder applied to $g(t) = \psi(x+tu)$ with $k = \lfloor s \rfloor$:

$$\begin{aligned}\psi(x+hu) =& \sum_{j=0}^{k-1} \frac{h^j}{j!} g^{(j)}(0) + \int_0^h g^{(k)}(t) \frac{(h-t)^{k-1}}{(n-1)!} dt \\ =& \sum_{j=0}^{k} \frac{h^j}{j!} g^{(j)}(0) + \int_0^h (g^{(k)}(t) - g^{(k)}(0)) \frac{(h-t)^{k-1}}{(n-1)!} dt.\end{aligned}$$

The first term is $\psi(x)$ plus a polynomial in u which will vanish after insertion in (31) because K is orthogonal the first non-constant polynomial of degree $\leq r$. The second term is bounded as

$$|\int_0^h (g^{(k)}(t) - g^{(k)}(0)) \frac{(h-t)^{k-1}}{(k-1)!} dt| \leq C|u|^k h^{k-1} \int_0^h \|\psi^{(k)}(x+tu) - \psi^{(k)}(x)\| dt.$$

Hence

$$|\int (\psi(x+hu) - \psi(x)) K(u) du| \leq C h^{k-1} \int_0^h \int \|\psi^{(k)}(x+tu) - \psi^{(k)}(x)\| |u|^k K(u) du\, dt \tag{32}$$

and by the generalized Minkowski inequality [25][2]

$$\begin{aligned}\|\psi_h(X_1) - \psi(X_1)\|_2 \leq& C h^{k-1} \int \left(\int \|\psi^{(k)}(x+tu) - \psi^{(k)}(x)\|^2 u^{2k} K(u)^2 1_{0\leq t\leq h} f(x) dx \right)^{1/2} du dt \\ \leq& C' h^{k-1} \int \left(|tu|^{2\alpha} |u|^{2k} K(u)^2 \right)^{1/2} 1_{0\leq t\leq h} du dt \\ \leq& C' h^{k+\alpha}.\end{aligned}$$

This proves (30). Concerning (29), we use (32) with f and $k = r$:

$$\begin{aligned}|f_h(x) - f(x)| \leq& C h^{r-1} \int_0^h \int \|f^{(r)}(x+tu)\| |u|^r K(u) du\, dt \\ \leq& C'' h^r \int |u|^r K(u) du\, ds\end{aligned}$$

□

Theorem 5. *(Efron-Stein inequality) Let $X_1, \dots X_n$ be an i.i.d. sequence, X_1' be an independent copy of X_1 and f be a symmetric function of n variables, then*

$$Var(f(X_1, \dots X_n)) \leq \frac{n}{2} \mathbb{E}[(f(X_1, \dots X_n) - f(X_1', X_2, \dots X_n))^2].$$

[2]For any nonegative function $g(.,.)$ on $\mathbb{R}^{k+p}$,

$$\left(\int \left(\int g(y,x) dy \right)^2 dx \right)^{1/2} \leq \int \left(\int g(y,x)^2 dx \right)^{1/2} dy$$

Theorem 6. *If the support of φ is a bounded convex set and φ is α-Hölder inside its support then $\varphi \in \mathcal{H}_{\min(\alpha,1/2)}$.*

Proof. We have

$$\begin{aligned}\int |\varphi(x+u)-\varphi(x)|^2 dx &\leq \|\varphi\|_\infty^2 \int (1_{\{x+u\in Q\}}1_{\{x\notin Q\}} + 1_{\{x+u\notin Q\}}1_{\{x\in Q\}})dx + C'|u|^{2\alpha} \\ &\leq \|\varphi\|_\infty^2 \lambda(y \,:\, dist(y,\partial Q) < |u|) + C'|u|^{2\alpha} \\ &\leq \|\varphi\|_\infty^2 \xi_{n-1}(S)|u| + C'|u|^{2\alpha},\end{aligned}$$

where $\xi_{n-1}(S)$ is called a Quermassintegrale of Minkowski. The last inequality follows from the Steiner's formula stated for instance in [8], Theorem 3.2.35 page 271. □

The following lemma gives some bounds on the conditional moments of ξ_{12} that are useful in the proof of Theorem 1.

Lemma 7. *Let $\xi_{ij} = K_{ij} - f_h(X_i)$, under (B1) and (B2)*

$$\mathbb{E}[\xi_{12}|X_1] = 0 \tag{33}$$

$$\mathbb{E}[|\xi_{12}|^p|X_1] \leq Ch^{-(p-1)d} \tag{34}$$

$$|\mathbb{E}[\xi_{13}\xi_{23}|X_1,X_2]| \leq \|f\|_\infty(h^{-d}K_2(h^{-1}(X_2-X_1)) + \|f\|_\infty), \tag{35}$$

with $K_2(x) = \int |K(x-y)K(y)|dy$.

Proof. The first equation is trivial. For the second equation, the triangular inequality and the Jensen inequality provide

$$\mathbb{E}[|\xi_{12}|^p|X_1]^{1/p} \leq 2\mathbb{E}[|K_{12}|^p|X_1] = 2h^{-(p-1)d}\int |K(u)|^p f(X_1+hu)dx,$$

and the third one is derived by

$$\begin{aligned}|\mathbb{E}[\xi_{13}\xi_{23}|X_1,X_2]| =& |\mathbb{E}[\xi_{13}K_{23}|X_1,X_2]| \\ =& h^{-d}|\int (h^{-d}K(h^{-1}(x-X_1)) - f_h(X_1))K(h^{-1}(x-X_2))f(x)dx| \\ =& |\int (h^{-d}K(h^{-1}(X_2-X_1)+u) - f_h(X_1))K(u)f(X_2+hu)dx| \\ \leq& \|f\|_\infty(h^{-d}K_2(h^{-1}(X_2-X_1)) + \|f\|_\infty).\end{aligned}$$

□

Acknowledgement. The authors would like to thank Céline Vial for helpful comments and advices on this article.

References

[1] Russel E Caflisch. Monte carlo and quasi-monte carlo methods. *Acta numerica*, 1998:1–49, 1998.

[2] Song Xi Chen. Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2):131–145, 1999.

[3] R Dennis Cook and Sanford Weisberg. Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

[4] Arnak S. Dalalyan, Anatoly Juditsky, and Vladimir Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *J. Mach. Learn. Res.*, 9:1648–1678, 2008.

[5] Michel Delecroix, Marian Hristache, and Valentin Patilea. On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference*, 136(3):730–769, 2006.

[6] LP Devroye and TJ Wagner. The strong uniform consistency of kernel density estimates. *Multivariate analysis*, 5:59–77, 1980.

[7] J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1996.

[8] Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[9] Fabrice Gamboa, Jean-Michel Loubes, and Elie Maza. Semi-parametric estimation of shifts. *Electron. J. Stat.*, 1:616–640, 2007.

[10] Theo Gasser and Hans-Georg Müller. *Kernel estimation of regression functions*. Springer, 1979.

[11] P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980. Probability and Mathematical Statistics.

[12] Wolfgang Härdle. *Applied nonparametric regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 1990.

[13] Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.*, 84(408):986–995, 1989.

[14] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, 29(3):595–623, 2001.

[15] Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993.

[16] MC Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

[17] Ker-Chau Li. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86(414):316–342, 1991.

[18] François Portier and Bernard Delyon. Optimal transformation: a new approach for covering the central subspace. *J. Multivariate Anal.*, 115:84–107, 2013.

[19] François Portier and Bernard Delyon. Bootstrap testing of the rank of a matrix via least squared constrained estimation. *J. Amer. Statist. Assoc.*, 2014.

[20] James L. Powell, James H. Stock, and Thomas M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989.

[21] Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954, 1988.

[22] L Rutkowski. On-line identification of time-varying systems by nonparametric techniques. *Automatic Control, IEEE Transactions on*, 27(1):228–230, 1982.

[23] Charles J Stone. Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, pages 267–284, 1975.

[24] Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

[25] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, 2009.

[26] Céline Vial. *Deux contributions à l'étude semi-paramétrique d'un modèle de régression*. PhD thesis, University of Rennes 1, 2003.

[27] Yingcun Xia. A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.*, 35(6):2654–2690, 2007.

[28] Peng Zeng and Yu Zhu. An integral transform method for estimating the central mean and central subspaces. *J. Multivariate Anal.*, 101(1):271–290, 2010.